디지탈 음성신호처리

이 황 수
(한국과학기술원 전기 및
전자 공학과 조교수)

1. 머릿말

음성은 인간과 인간사이의 가장 자연스러운 의사 전달, 즉 통신수단이다. 따라서 어떻게하면 효율적으로 멀리 떨어져 있는 두 지점 사이의 음성통신을 할 수 있느냐가 전화가 발명된 이래 음성신호 처리에 있어서 주된 관심사로 연구되고 있다.이러한 연구는 최근들어 급속한 발전을 보이고 있는 반도체 및 컴퓨터기술에 힘입어 양자화된 음성신호, 즉 디지탈음성신호 처리 분야에 집중되고 있으며,인간과 인간사이만이 아닌 인간과 컴퓨터, 다시말하면 인간과 기계사이의 통신에서도 음성을 이용하려는 방향으로 진전되고 있다.

음성신호의 처리능력을 규정하는 데에는 여러가지 방법이 있을수 있다. 우선, 첫째로 정보이론적인 면에서 보는 정량적인 방법을 들 수 있다. 정보이론에 의하면 음성신호는 그 음성이 갖고있는 메시지의 내용, 즉 정보로 나타낼 수 있다. 또, 다른 방법으로는 메시지정보를 포함하고 있는 신호, 즉 음향신호의 파형으로 규정하는 수도 있다. 이들중 복잡한 통신시스템을 해석하는 데에는 정보이론적인 개념이 중요한 역할을 하게 되지만 실제적인 음성통신의 응용분야에 있어서는 음성을 파형이나 파라메터를 이용한 특정한 모델로 생각하는 것이 더 유용하다.

음성통신의 과정을 간략히 살펴보면 다음과 같다. 말하고자 하는 사람의 머리에 관념적인 형태로 떠오른 메시지로부터 시작하여 음성발생의 복잡한 과정을 거치면서 메시지가 나타내는 정보가 궁극적으로는 음향신호로 바뀌게 된다. 메시지정보들은 이와같은 음성발생 과정을 거쳐 각기 다른 형태로 표현된다고 생각할 수 있다. 예를들면, 각 메시지정보들은 일단 입술, 혀, 성도 등 음성발생 기관을제어하는 신경신호들로 바뀌게 되고, 각 기관들이 신경신호에 따라 일련의 동작을 함으로써 그 결과로 원래의 관념을 나타내는 메시지정보를 음향파형으로 나타낸다.

음성을 통하여 교환하려는 정보는 각각 분리된 형태로 되어 있다. 즉, 이 정보는 유한한 심볼의 집합에서 뽑은 요소들이 연결된 것으로 볼 수 있다. 서로다른 소리를 구별지어주는 이 각각의 심볼을 음소라 한다. 인간이 사용하는 특정한 언어마다 고유한 음소를 갖고 있으며, 그 숫자는 대략 30개에서 50개 사이이다.

정보이론에서의 주된 관심사는 정보가 어떠한 속도로 전달되느냐에 있다. 음성에 있어서는 발성기관들이 실제적으로 움직이는 속도에 제한이 있기때문에 이 정보전달속도를 대략 추정하여 볼 수 있는데 1초당 약 10개의 음소가 발음된다고 할 수 있다. 만약에 각각의 음소들을 2진수로 나타낸다고 하면 음소 전체를 부호화하는데 6비트($2^6 = 64$가지)이면 충분하다. 평균적으로 1초당 10개의 음소를 발음한다고 가정했을 경우에 연속되는 음소간의 상관관계를 무시하면 음성의 평균정보전달 속도는

60 bps (bits per sec) 정도가 된다. 다시말하면, 정상적인 속도로 발음할때 문자로 표시할 수 있는 음성의 정보량은 60bps정도가 된다. 물론 실제로 음성의 정보량은 문자로 표시될 수 있는 양 이외에 말하는 사람에 따른 정보, 말하는 사람의 감정 상태, 말하는 속도 및 음량의 변화를 포함할 경우에는 이보다 훨씬 더 높아지게 된다.

음성통신시스템은 음성신호를 여러가지 방법으로 전송, 저장, 처리하게 된다. 이와같은 일을 기술적으로 해결하기 위하여 음성신호를 여러가지 형태로 표시하고 처리하게 되는데, 이때 일반적으로 다음의 두가지 점이 특히 문제가 된다. 첫째, 음성신호의 메시지내용을 보존하여야 하는 일이고 둘째 메시지 내용을 심하게 손상시킴이 없이 전송과 저장에 편리하도록 음성신호를 적절한 형태로 다르게 표현하는 것이다. 음성신호를 다른 형태로 바꿀때 유의할 점은 듣는 사람이 쉽게 내용을 알아들을 수 있거나 기계가 자동적으로 그 내용을 알 수 있도록 해야한다. 이와같이 음성신호의 형태를 바꾸어 나타낼 경우에 이를 표시하기 위해서는 500bps이상이 필요하다.

음성신호를 다른 형태로 표현하는 음성부호화의 경우, 되도록 그 표현에 필요한 비트의 수를 줄이는 것이 음성의 전송 또는 저장의 측면에서 바람직하다. 음성신호를 되도록 적은 비트수로 표현하려는 시도는 1939년 Dudley가 뉴욕의 세계박람회에서 선보인 음성부호화기 (Voder)로부터 시작된다. 이 음성부호화기는 인간의 음성을 흉내낸 듯한 소리를 내도록 되어있어 비록 그 소리를 알아들을 수는 있어도 그리 음질이 좋지는 못하였다. 따라서 사람들은 과연 이것을 실제로 사용할 수 있을까하고 생각하였다.

그러나, 그때와는 비교할 수 없을 정도로 과학이 발달된 오늘날에 와서도 음성부호화기의 구조는 훨씬 더 복잡하여졌으나 그 음질은 전송속도가 9.6 Kbps이하가 될 경우 Dudley의 음성부호화기보다 별로 나아지지 않고 있다. 음성을 기계로 인식하려는 연구도 현재까지는 적절한 비용으로 기계와 자연스러운 대화를 사용하여 의사소통을 하는 데까지는 미치지 못하고 있다. 어떠한 제한된 형태로 발음된 말을 알아들을 수 있는 기계들은 만들어지고 있지만 이것이 어떤 지능을 갖고 있다고 말할 수는 없다. 사람이 구술하는 내용을 알아듣고 글로 바꾸어줄 수 있는 타이프라이터도 아직 완성되지 않고 있다. 그러나, 음성처리 기술의 진보로 인하여 인쇄 회로기판 한장으로 음성을 알아듣고 그에따라 기계를 제어할 수 있도록 되어있는 제품이 저렴한 가격으로 제공되기 시작하고 있으며, 비교적 느린 음성인식 분야의 기술진보 속도에도 불구하고 이의 실용화 노력은 매우 활발히 진행되고 있다.

2. 음성부호화

연속적인 음성신호를 어떻게 부호화하여 디지탈 형태로 변환시키고, 음성전송 회선과 저장장치를 효율적으로 사용할 수 있도록 하는 음성의 부호화기술에 대하여 살펴보자.

우선 생각하여 볼 수 있는 것으로 아날로그방식이든 디지탈방식이든 간에 현재 한 채널만 사용하던 통신선로에 둘 또는 그 이상의 음성채널의 통화를 가능케 할 수 있다면 새로운 통신선로의 설치에 필요한 많은 노력과 비용을 절약할 수 있을 것이다. 나아가 이와같은 한 통신선로로 음성과 데이터 신호를 공동으로 전송할 수 있다면 새로운 서어비스를 더 빠르고 경제적으로 제공할 수 있게 된다.

디지탈음성 저장시스템의 경우는 될수록 낮은 전송속도에서 동작하는 음성부호화방식을 사용할 때 기억장치와 대역폭을 경제적으로 사용할 수 있게된다. 최근들어 쓰이기 시작하고 있는 광섬유 선로를 쓸 경우 매우 넓은 대역폭을 사용할 수 있으므로 음성전송에서의 전송속도감축이 과연 아직도 필요한가에 대한 의문을 제기할 수도 있다. 앞으로 전 통신선로가 광섬유로 대치된다면 그와같은 일이 필요없을 수도 있겠으나 여하튼 대역폭이 제한된 통신채널이 존재하는한 전송속도 감축을 위한 음성부호화방식은 계속 개발되어야할 것이다.

음성부호화방식에 관한 연구도 많이 발표되어 쓰이고 있지만 여기에서는 PCM (pulse code modulation), 적응차등 PCM (ADPCM : adaptive differential PCM), SBC (subband coding)방식과 선형예측부호화 (LPC : linear predictive coding)방식에 대하여 간략히 설명하고자 한다.

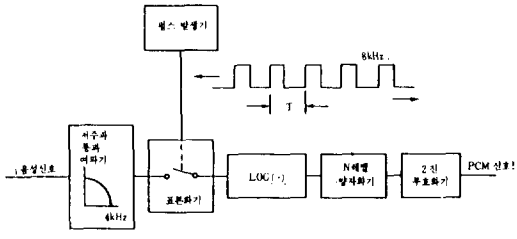

그림 1. PCM부호화과정

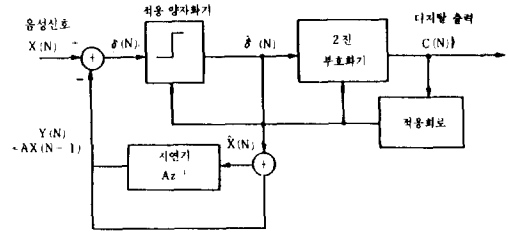

그림 2. ADPCM부호화과정

2.1 PCM방식

현재 우리나라에서 전화선을 이용한 음성의 디지탈통신에 주로 쓰이고 있는 PCM방식은 μ-law PCM으로, 그 부호화방식은 그림 1에 나타나 있다. 먼저 연속된 음성신호가 들어오면 저주파통과여파기(low pass filter)를 통과시켜 대역폭을 제한한 다음, 표본화주파수 8 KHz와 8비트(2^8=256레벨)디지탈신호로 바꾸게된다. 이경우 연속된 값을 갖는 입력음성신호를 256개의 값을 갖는 디지탈신호로 변환시키는 과정에서 필연적으로 잡음이 발생하게 되는데 이를 양자화잡음(quantization noise)이라 한다. 따라서 음성부호화의 과정에서 어떻게 하면 이 양자화잡음을 줄일수 있는가가 문제가 된다. μ-law PCM의 경우는 음성파형의 크기가 대수함수(logarithmic function)적으로 분포되어 있는 확률적 성질을 이용하여 이 잡음을 줄여 신호 대 잡음의 비율을 높이고 또한 입력신호의 크기에 관계없이 잡음을 일정한 레벨로 유지시켜 전화를 듣는 사람으로 하여금 잡음을 느낄 수 없도록 하여주고 있다. 그림 1에서 보는 바와 같이 우선 표본화된 음성신호의 대수함수를 취하여 큰 음성신호는 그 크기를 줄이고 작은 음성신호는 그 크기를 키워준 다음 등간격으로 나뉘어져있는 8비트(256 레벨) 양자화기를 거쳐 전송하기에 편리하도록 부호화하게 된다. 그러므로 현재 전화선로에서 사용하고 있는 μ-law PCM의 경우 64 K bps의 전송속도를 갖게된다. 이경우 이 디지탈 신호를 전송하기 위한 통신채널의 대역폭은 입력아날로그신호를 그대로 전송할 때 필요한 4KHz보다 더 넓게 된다. 이와같이 통신선로의 대역폭을 더 차지하면서도 입력아날로그 음성신호를 디지탈신호로 바꿔보내는 이유는, 디지탈신호는 2진수로 표시되어 있기 때문에 통신선로가 길어짐에 따라

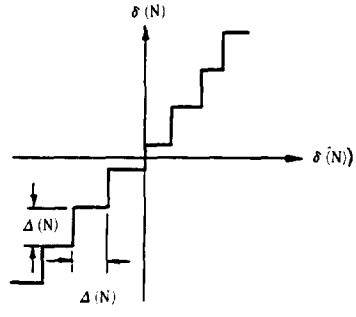

그림 3. ADPCM의 3비트 양자화기

누적되어 발생하는 잡음의 영향을 줄일수가 있고,따라서 통신선로의 중간에 설치하는 중계기의 수도 아날로그음성신호를 그대로 보낼때에 비하여 대폭적으로 줄일수가 있기 때문이다. 이 μ-law PCM방식에 의하여 부호화된 음성신호의 음질은 귀로 들어서는 원래의 음성신호와 구별할 수 없을 정도이다.

2.2 ADPCM 방식

ADPCM 방식은 표본화된 음성신호의 인접신호들 사이의 상관관계를 이용한 부호화방식으로 음질은 유지하면서 그 전송속도를 PCM 방식에 비하여 두배이상 낮출수 있다. 현재 표준화되어 사용하기 시작한 ADPCM의 경우 32Kbps의 전송속도로 64K 비트 PCM방식에 비하여 통신채널을 두배 효율적으로 사용하면서도 그 음질은 PCM 방식과 거의 동일하다. 간단한 ADPCM 방식의 부호화과정이 그림 2에 표시되어 있다. 그림에서 보면 입력신호 X(n)과 입력신호의 추정치 Y(n)의 차이 δ(n)이 적응양자화기에 의하여 디지탈화되어 전송되도록 되어있다. 적응양자화기는 δ(n)이 양자화된 출력디지탈신호의 패턴에 따라 양자화기의 각 레벨간의 간격

을 조정하여주게 되어있다. 이때 실제의 음성신호와 같이 인접한 표본신호사이에 상관관계가 있다면 X(n)과 Y(n)의 차신호인 δ(n)은 X(n)보다 매우 작은 신호가 될 것이고 따라서 이 양자화잡음도 그에따라 줄일수가 있게된다. 양자화기의 기능은 그림 3에 보인 3비트(8레벨) 양자화기로 설명할수 있다. 만약 입력신호가 양자화기의 바깥쪽 레벨로 부호화되었다면 입력음성신호의 상관관계에 의하여 입력신호들이 커질 것이라는 것을 알 수 있다. 따라서 이 경우는 각 양자화레벨사이의 간격 Δ(n)을 늘려주어 양자화기가 큰 입력신호도 양자화할수 있도록 하여준다. 물론 입력신호가 양자화기의 안쪽 레벨로 부호화되었을 경우는 이와 반대로 Δ(n) 을 줄여주게 된다. 이와같은 양자화기를 적응양자기라 하고 4비트(16레벨) 양자화기를 쓸 경우 32K bps의 전송속도를 필요로 한다.

2.3 SBC 방식

앞에서 설명한 ADPCM의 경우 32K bps의 전송속도로 64Kbps PCM과 거의 같은 성능을 낼수 있는데, 이보다 더욱 전송속도를 줄인 음성부호화방식들이 개발되어 있다. 그중의 하나로 subband 부호화방식을 들 수 있다. 이 방식은 음성을 각각 다른 주파수대역을 갖는 여러개의 주파수 대역여파기를 통과시켜 각 주파수대역마다를 ADPCM으로 부호화하게 된다. 각 대역에 대한 전송속도는 그 대역의 음성청취에 대한 기여도에 따라 다르며, 이로 인하여 전체의 전송속도가 16K bps이하에서도 상당히 좋은 성능을 나타낸다. 실제로 음성부호화기들의 성능을 직접 청취하여 시험하여 보면 전송속도가 16K bps이하에서 급격한 성능의 저하를 발견하게 된다. 따라서 보통 전송속도가 9.6K bps 이하인 부호화방식을 사용한 음성신호를 저전송속도 음성신호(LBRV : low bit-rate voice)라 한다.

2.4 선형 예측 부호화 방식

음성의 전송속도를 더욱 낮추기 위하여는 주로 선형예측부호화(LPC)방식을 이용한 음성부호화기를 사용하게 된다. 최근 10여년에 걸쳐 LPC 방식을 이용하면 전송속도를 1.2K비트 이하로 내리면서도 자연스러운 원래 음성의 음질을 얻을 수 있을 것으로 생각되어왔다. 이와같이 낮은 전송속도에서 음질이 만족스러운 부호화방식을 아직까지는 연구해내지 못하고 있으나 그 유용성으로 인하여 서서히 실용화 할 수 있는 제품들을 선보이고 있다. LPC 방식을 간략히 요약하여 보면 다음과 같다.

LPC방식은 음성파형 자체를 다루는 앞서의 부호화방식들과는 다른 원리를 이용하고 있다. 즉 각기 다른 음성이 성도(성대에서부터 입 또는 코로 통하는 공간)의 주파수 특성을 변화시켜 발생되는 점을 근사화시켜 부호화에 이용하고 있다. 이 LPC방식에서는 사람의 성도를 시간에 따라 변화하는 계수(선형예측계수 또는 LPC계수라 함)를 갖는 선형여파기로 모델화하고 있다. 이 여파기를 구동시키는 입력신호로는 유성음발생의 경우 성대의 진동을 준주기적인 펄스로, 무성음발생의 경우는 성도의 수축, 마찰로 인한 공기의 흐름을 백색잡음으로 모델화하여 사용한다. 따라서 각각의 유성음, 무성음에 따라 입력신호의 크기 또는 주기 등이 변하게 된다. 이와같은 음성에 대한 각종계수들을 구하는 과정을 LPC 해석이라 하는데 보통 자기상관법(auto-correlation method)을 사용하며, 대략 10개 정도의 LPC계수를 구하여 선형여파기의 계수로 사용한다.

LPC부호화기로 LPC해석에서 구한 LPC계수, 음성의 주기, 입력신호의 크기와 종류에 대한 정보를 부호화하여 전송하게 되는데 수신측에서는 이와 같은 정보들을 받아서 다시 음성을 합성하게 된다. 이를 LPC합성이라 하며 모델화된 선형여파기의 계수로 LPC계수를 사용하고 입력신호로는 수신된 준주기적 펄스신호나 백색잡음 신호를 사용하여 음성을 합성해서 수신자에게 들려주게 된다.

3. 음성부호화의 응용

음성부호화의 새로운 방식들의 개발과 각종 음성신호처리용 소자들의 출현으로 음성부호화 방식이 여러분야에 응용되고 있다. 이중 특히 괄목할만한 것은 대규모 집적회로(VLSI : very large scale integration)기술의 진보로 많은 음성부호화 제품들에 실용화되고 있다.

불과 수년 전만 하더라도 LPC부호화기의 생산은 일반적인 디지탈 집적회로인 TTL IC가 쓰였고 그

디자인에도 특수한 기법이 요구되었다. 또한 이와 같이 제작한 부호화기는 우선 그 부피가 클뿐만 아니라 전력손실도 크고 또한 음질도 만족할만하지 못하여 널리 쓰이지 않았다. 그러나, 이러한 상황은 디지탈 신호처리(DSP : digital signal processing) IC의 개발로 크게 달라지게 되었다. 새로운 DSP IC들과 고급 언어를 사용한 이의 개발 장비들의 등장으로 이제는 프로그램만 적절히 개발함으로써 각종 부호화기들을 값싸게 만들수 있게되어 많은 응용분야에 이들 부호화기들을 사용할 수 있게 되었다. 이중 몇가지 대표적인 응용분야를 살펴보면 다음과 같다.

3.1 무선전화

무선전화에서 사용하는 주파수대역이 제한되어 있는 관계로 음성의 전송속도를 매우 낮춰야만 많은 가입자를 수용할 수 있다. 따라서 이 경우에 낮은 전송속도의 음성부호화기 일수록 유리하며 또한 디지탈 신호의 전송이기 때문에 전화내용의 기밀유지에 필요한 음성의 암호화를 쉽게 이룰 수 있다.

3.2 음성 저장 및 전달

대표적인 것으로 다중채널을 갖는 디지탈 전화응답기를 들 수 있다. 이 전화응답기는 많은 부가기능을 가질수 있는데 예를들면, 가입자의 음성 메시지를 저장하였다가 가입자가 원하는 날짜와 시간에 이 메시지를 다른 가입자에게 전달할 수가 있다.또한 이 시스템을 이용할 경우 일절한 시간에 상대방에게 전화를 자동적으로 걸수 할수 있을 뿐만 아니라 가입자들끼리의 인사말도 쉽게 나눌수가 있다. 이때에도 낮은 전송속도의 음성부호화기를 사용하면 시스템의 기억용량과 대역폭을 효율적으로 사용할 수 있다. 우리나라에서도 금년도에 시범으로 이와같은 서어비스를 실시할 예정이다.

3.3 컴퓨터 음성응답

이는 전화선을 이용하여 여러가지 정보를 얻는 컴퓨터를 이용한 시스템에 응용될 수가 있는데 그 응용분야는 매우 넓다. 예를들면 전화번호 자동안내, 시간안내, 각종 여행정보안내, 날씨안내 등을 들수가 있겠는데, 이 시스템들의 경우는 기본적인 단어나 문장들이 부호화되어 저장되어 있다가 프로그램이 지시하는 대로 적절히 합성되어 음성메시지를 만들어 내는 일을 한다.

3.4 자동 전화교환 시스템

일반적인 사무자동화 추세의 진전으로 전화선을 이용하여 많은 양의 데이타(특히 컴퓨터 데이타)를 전달하려는 요구가 커지고 있다. 이경우 음성을 디지탈로 부호화하면 데이타와 같이 동일한 통신회선을 이용할 수가 있고, 또한 개인의 비밀유지를 위한 암호화도 쉬워져 통신시설을 좀더 경제적으로 사용할 수가 있게 된다.

3.5 사설통신망

앞의 자동 교화교환 시스템과 동일한 이점을 갖고있다. 특히 국제전화의 경우 음성부호화기를 사용하여 한 회선으로 여러명이 동시에 통화를 하거나 통화시간을 단축하여 비용을 크게 줄일 수 있다.

3.6 데이타통신망을 통한 음성통신

패켓(packet) 교환방식을 이용하는 데이타통신망에 패켓으로 만든 음성신호도 전송함으로써 전송로의 이용율을 높일 수 있다. 우리나라에도 현재 패켓 데이타통신망이 설치되어 있어 이 통신망을 이용한 음성통신방식도 개발, 연구되고 있다. 그러나 데이타와 음성의 서로 다른 특성 때문에 새로운 통신방식의 개발이 필요하다. 현재의 통신망에서는 전송하려는 데이타의 양이 통신망의 수용능력을 넘으면 보통 통신을 할 수 없게 되어 있다. 그러나, 음성의 패켓통신의 경우는 음성부호화의 전송속도를 적절히 조절하여 약간 음질은 떨어지지만 통화가 가능하도록 할 수 있다.

이상으로 음성부호화의 많은 응용분야중 현재 응용이 가장 활발히 진행중인 몇분야를 살펴보았다.

4. 자동 음성인식

음성인식에 대한 연구는 60년대 후반부터 본격적으로 시작되어 70년대를 거치면서 계속되어 왔으나 아직도 궁극적인 목표 즉, 어느 사람의 말이건 간에 어휘의 제한없이 연속음성을 알아듣는 능력을

갖는 인식방법의 연구에는 훨씬 못미치고 있다. 이것은 현재 활발히 진행중인 인공지능 연구가 결실을 맺기 전까지는 가능하지 못하리라 생각된다.

현실적으로는 특정한 사람이 아닌 여러사람이 연속적으로 발음하는 음성을 알아들는 인식시스템도 그 어휘의 수가 100단어 정도가 넘는 것은 만들기가 어렵다. 이는 사용하는 언어에 따라서도 다르다. 일본어의 경우는 영어의 경우보다 철자법이나 발음의 모호성이 적기 때문에 80년대 말경에는 특정인이 아닌 여러사람이 발음하는 연결언어를 어휘의 제한없이 인식하는 시스템의 개발이 가능하리라 예견된다. 우리말의 경우도 영어에 비하면 역시 이와 같은 이점이 크기때문에 연구가 집중된다면 빠른 시일내에 좋은 결과를 낼 수 있을 것이다. 특히, 일본어의 경우는 1000여 자가 넘는 자판을 필요로 하는 타이프라이터의 대체를 위하여도 자동 음성인식 시스템의 연구개발에 큰 힘을 쏟아넣고 있다.

현재의 음성인식 시스템이 어떠한 상태까지 연구되어 있는가를 살펴보자. 우선 특정한 사람의 음성을 인식하는 시스템은 격리단어인식(isolated word recognition)의 경우 상당히 많은 양의 어휘를 인식할 수 있다. 그러나, 이때에도 어휘수에는 제한이 생기게 되는데 이는 주로 많은 어휘에 대한 각 패턴의 혼동과 말하려는 사람에게 맞는 인식시스템의 훈련이 필요하기 때문이다. 여러사람의 음성을 인식하는 시스템에서는 격리어휘의 경우에도 적은 양의 단어 밖에 인식할 수 없는데 그것도 이 인식시스템을 사용하려는 집단의 구성원에 맞도록 통계적인 훈련과정을 거쳐야 하고 또한 각 단어들을 정확하게 인식하려면 제한된 인식 어휘를 선택해야 하며, 그 음이 크게 다른 단어들을 택할 필요가 생기게 된다. 예를 들면 숫자를 나타낼 때 일, 이, 삼, 사,… 대신에 하나, 둘, 셋, 넷,… 등을 택한다던가 하는 일인데, 이는 일반적인 생활습관 때문에 쉬운 일은 아니다. 이렇게 어휘를 선택한 후 주의하여 발음을 하더라도 인식시스템의 인식율은 그리 높지 않은 것이 보통이다. 사람의 언어전달의 경우는 기계와 달리 어떤 특정한 단어의 발음이 불명확하여도 전후 문맥이나 생활습관을 통하여 또는 재확인을 통하여 알아들을 수가 있지만, 현재의 컴퓨터를 이용한 음성인식 시스템은 문장의 문법규칙이나 재확인 과정만을 이용하고 있어서 응용분야에 따라서는 그 사용이 부적절한 경우가 많다.

자동 음성인식 시스템의 인식과정을 여기에서는 격리단어 인식에 대하여 설명하기로 한다. 격리단어 인식시스템은 지금 현재 상당한 기술적 진보를 이루어 실용화 단계에 와있고, 그 대부분의 구성요소들은 연결단어 인식시스템(connected word recognition)에서도 동일하게 사용할 수 있다. 격리단어 인식시스템은 보통 단어 또는 구를 인식하게 되는데 대개 실제의 응용분야에서 필요로하는 어휘의 길이보다는 짧은 것이 제약점이다.

음성인식 시스템은 발음의 제약조건에 따라 격리단어 인식시스템, 연결단어 인식시스템, 연속 음성인식시스템(continuous speech recognition)으로 나눌 수 있다. 격리단어 인식시스템은 그 명칭이 나타내듯이 단어와 단어사이에 적은 묵음 구간을 필요로 한다. 반면 연결단어 인식시스템은 각각의 단어만 정확히 발음하면 된다. 연속음성 인식시스템은 일반 대화를 하듯이 발음하는 음성을 인식하는 시스템으로써 어휘들이 연속적으로 발음되며, 또한 발음하는 사람에 따라 달라지는 변이음현상으로 인하여 음성인식 작업이 한층 더 어려워진다.

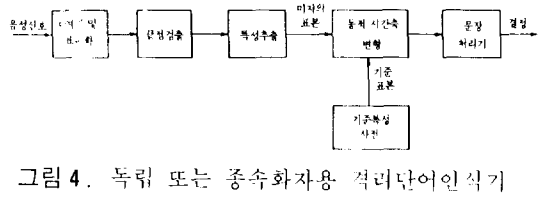

그림 4. 독립 또는 종속화자용 격리단어인식기

격리단어 인식시스템의 구성을 그림 4에 나타내고 있다. 이 그림에서 보는 인식시스템의 구성도는 다른 인식시스템에서도 거의 동일하다. 맨 첫단에 있는 여파기와 표본화기는 입력음성신호의 주파수 특성을 인식시스템에 맞도록 조정한 다음 입력신호의 주파수대역에 맞는 표본화율로 아날로그 음성신호를 디지탈화 한다.

다음 블록은 음성의 시작과 끝점 검출부분이다. 여기에서는 음성의 시작과 끝점 검출뿐만 아니라 음성의 유·무성, 비음의 여부와 주위의 잡음의 영향을 제거하는 역할을 한다. 이를 위하여 음성의 에너지와 영교차율등을 위시한 여러가지 파라메터와

음성이 입력되는 주위 환경에 대한 정보를 복합적으로 이용하게 되는데 여기에는 주로 경험적인 수치들이 사용된다.

입력음성의 특성추출에서는 음성의 특성 파라메타로 각 주파수대역에서의 에너지 또는 LPC 계수들을 추출하게 된다. LPC계수를 음성인식의 특성 파라메타로 사용하는 방법이 가장 보편적이며 그 정확도도 높은 것으로 알려져 있다. 이 LPC 계수는 음성의 주파수특성을 잘 나타내어 유사한 음성의패턴에 잘 맞음은 물론 계수의 수가 입력음성 자체에 비하여 상당히 적으므로 다루기가 용이하다는 이점이 있다. 음성의 특성 추출은 일정구간의 입력음성 블록에 대하여 하게 되는데, 그 구간은 보통 10-20ms가 된다.

일단 입력음성의 특성 파라메타가 추출되면 이것이 음성 인식 시스템에서 보유하고 있는 어떤 단어의 기준 패턴과 가장 유사한가를 결정하게 된다.패턴의 유사도 결정에서 가장 문제가 되는 것은 같은 단어의 발음에 있어서도 그 발음시간에 차이가 있는 점이다. 이것은 서로 다른 음성에서는 물론 한 사람의 발음에서도 발음할 때마다 조금씩 다르게 되는 것을 볼 수 있다. 따라서 이와 같은 시간축 상에서의 단어의 발음길이를 조정하기 위하여 시간축 변형을 하게 되는데 비교하려는 기준 단어의 패턴에 맞추어 동적으로 행한다. 이를 동적 시간축 변형 (DTW : dynamic time warping)이라 하며 변형된 시간축에 따라 패턴 유사도 측정을 모든 기준 단어들에 대하여 행한 다음, 그 유사도가 가장 큰 단어를 선택하게 된다. 인식의 정확도는 보통 특정 화자에 대하여 99% 이상을 얻을 수 있으며 어휘수가 적은 여러사람의 음성을 인식할 수 있는 시스템의 경우에는 98% 이상의 정확도를 얻고 있다.

음성인식 시스템이 어느 특정한 한 사람의 음성을 인식하느냐 또는 여러사람의 음성을 인식하는가는 인식시스템이 갖고 있는 기준패턴사전을 어떻게 만드느냐에 달려있다. 특정화자의 음성을 인식할 경우 이 기준패턴사전은 그 사람의 음성의 특성을 추출하여 만들게 된다. 말하는 사람에 관계없이 음성을 인식하려 할 때에는 단어의 기준패턴사전을 여러개로 만들거나 또 여러사람의 발음에서 평균특성 패턴을 찾아내어 구성할 수가 있다. 이렇게 기준패턴사전을 구성하더라도 각 개인의 음성자체도 감기, 정서적인 요인 또는 발음하는 환경의 바뀜에 따라 달라지는 현상이 발생하므로 인식의 문제를 더욱 어렵게 하고 있다. 예를 들면 전투기 조종사의 경우 음성인식 시스템을 조용한 곳에서 훈련시킬때와 전투시에 발음할 때는 동일한 상태의 제어명령을 내리지는 못할 것이다. 이 경우 위의 환경도 시끄러운 상태로 바뀌지만 실제로 전투상황에서 겪는 스트레스로 인하여 목소리 자체도 달라지게 된다. 이점이 패턴의 유사도를 측정하여 음성인식을 행하는 시스템의 단점으로써, 인식시스템을 훈련시킬때와 다른 상황에서는 인식시스템의 성능이 저하된다. 이를 해결하기 위하여 인식시스템을 입력음성의 변화에 따라 시간이 지나면 적응되도록 하는 방식에 대한 연구도 행하여지고 있으나, 앞에서의 조종사의 경우와 같이 인식시스템의 적응에 대한 충분한 시간이 없을때에 얼마만큼 효과가 있을지는 의문시 된다.

말하는 사람에 관계없이 음성을 인식하려는 시스템의 경우는 시스템을 보통 통계적으로 많이 사용하게 될 사람들의 집단에 맞도록 훈련을 시키게 된다. 이때 한단어당 여러개의 기준패턴을 갖게 됨으로 인하여 인식할 수 있는 어휘의 수가 상대적으로 줄어들게 된다. 따라서 한 단어의 기준 패턴이 차지하는 기억용량이 크게 되며, 특히 중요한 것은 실시간으로 음성인식을 하기 위하여 패턴 유사도 결정에 소요되는 시간을 줄일 필요가 생기고 이로 인하여 전체적인 기준 패턴의 수, 즉 인식단어의 수가 줄어들게 된다. 많은 시스템들에서는 이를 해결하기 위해 약간의 구문규칙을 사용, 어느 한 시점에서 다음에 나타날 가능성이 있는 단어들의 수를 줄여 실시간 처리를 가능케 하고 있다.

5. 음성인식의 응용

음성인식시스템을 값싸게 구축할 수 있게되면 이를 이용한 많은 신제품과 향상된 서어비스를 기대할 수 있다. 많은 사람들이 음성인식 시스템을 실제로 이용하기를 원하고 있지만 현재까지의 음성인식 응용분야가 성공적이라고 할 수는 없다. 그 이유는 여러가지를 들 수 있겠으나 대개 그 응용분야가 적절치 않았거나, 응용분야에 대한 잘못된 이해, 그

리고 음성 입출력장치의 가격이 너무 높은 것이 문제였다.

앞에서도 언급하였듯이 이제 격리 단어 자동인식 연구는 실용화 단계에 와 있다. 그러나 그것은 최근의 일이며 이또한 인식방식의 연구개발과 더불어 반도체 소자의 개발에 힘입은 바가 크다. 범용 신호처리 VLSI기술 진보에 의해 음성인식제품의 실용화 개발이 경제적으로 가능해지고 있다.

최근들어 음성인식기술을 이용한 제품이 급격히 증가하고 있다. 그러나, 그 제품들이 과연 시장성이 있는지는 아직 의문이다. 인식인식시스템에서의 가장 큰 문제는 응용 소프트웨어이다. 시스템을 개발하기 전에 과연 이 시스템에 자동음성인식이 필요한가를 결정해야 한다. 단지 자동음성인식 기능을 사용하여 일을 해결할 수 있어서가 아니라(이 경우 보통 다른 수단으로도 일을 가능케할 수 있다.) 꼭 음성인식 기능이 있어야만 작업을 수행할 수 있는 것이라야 한다.

일단 어떠한 작업이 음성인식 기능을 통하여서만 최선으로 이루어질 수 있다고 결정되면 동작시키려는 소자나 시스템과 사람과의 대화를 최적화시킬 수 있는 응용 소프트웨어를 신중히 작성하여야 한다. 경제적인 가격으로 위와같은 음성인식시스템을 구성할 수 있다면 그 응용분야는 무궁무진하다고 할 수 있다. 음성으로 작동하는 각종 가정용 기기들을 비롯하여 전화를 이용한 24시간 정보 안내시스템등을 대표적인 예로 들 수 있다.

6. 맺음말

이상으로 디지탈 음성처리 분야의 대표적인 두 분야, 즉 음성부호화와 음성인식 분야의 전반적인 기술현황에 대하여 살펴보았다. 이외에도 음성처리 분야로 문자의 음성합성, 화자식별, 음성 이해분야 등을 생각할 수 있다. 그러나 이러한 분야들도 기본적으로는 여기에서 언급한 음성부호화와 인식분야의 연구결과들을 이용하고 있다. 이들이 복합된 최적의 응용분야로는 음성 자동번역시스템이 있다. 이 시스템의 완성은 인공지능에 대한 연구가 훨씬 더 진전되어야 하고, 이를 빠른 속도로 처리할 수 있는 제 5세대 컴퓨터가 구현되었을 때에 가능하다 하겠다.

참 고 문 헌

1) Engineering and Operations in the Bell System , Prepared by Members of the Technical Staff and the Technical Publication Department, Bell Laboratories, Indiana Publication Center, 1977, pp. 140-144.

2) P.Cummiskey, N.S.Jayant, and J.L.Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," Bell System Technical J., Vol. 52, No.7, Sept. 1973.

3) R.E.Crochiere, S.A.Webber, and J.L.Flanagan, "Digital Coding of Speech in Subbands," Int'l IEEE Conf.ASSP, 1976.

4) B.S.Atal and S.L.Hanauer, "Speech Analysis and Synthesis by Liner Prediction of the Speech Wave," J.Acoustics Soc. Amer., Vol. 50, No. 2, 1971, pp. 637-655.

5) J.Tierney, "A Study of LPC Analysis of Speech in Additive Noise,"IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-28, No. 4, Aug. 1980.

6) J.D.Markel and A.H.Gray, Linear Prediction of Speech, Springer-Verlag, New York, 1976.

7) L.R.Rabiner and R.W.Schafer, Digital Processing of Speech Signals, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1978, pp. 141-149.

8) B.S.Atal and J.R. Remde, "New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates," Proc. IEEE Conf. Trans. Acoustics, Speech and Signal Processing, Vol. 1, May 1982, pp. 641-617.

9) G.S.Kang and S.S. Everett, "Improvement of the Narrowband LPC Synthesis," IEEE Int'l Conf. Acoustics, Speech and Signal Processing, Vol. 1, No. 1, March 1984, pp. 1.7. 1-1.7.4.

10) The Bell System Technical J., Vol. 60, No. 7, Part 2, Sept. 1981.

11) G.White, "Speech Recognition : A Tutorial Overview," Computer, Vol. 9, No. 5, May 1976, pp. 40-53.

12) L.R.Rabiner and M.R.Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," Bell System Technical J.,Vol. 54, No. 2, Feb. 1975, pp. 297-315.

13) H.Sakoe and S. Chiba, "A Dynamic Programming Approach to Continuous Speech Recognition," Proc. 7th Int'l Cong. Acoustics, Budapest, Hungary, Aug. 1971, pp. 65-68.

14) F. Itakura,"Minimum Prediction Residual Applied to Speech Recognition," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, Feb. 1975, pp. 66-72.

15) J.L.Flanagan, "Computers that Talk and Listen : Man-Machine Communication by Voice," Proc. IEEE, Vol. 64, No. 4, April 1976, pp. 405-415.

16) J.Allen, "Speech Synthesis from Unrestricted Text," Speech Synthesis, J.L.Flanagan and L.R.Rabiner, eds., Dowden, Hutchinson, and Ross, Stroudsburg, Pa., 1973.

17) S.Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," IEEE trans. Acoustics, Speech and signal Processing, Vol. ASSP-29, No. 2, Apr. 1981, pp. 245-271.

18) A.E.Rosenberg and K.L,Shipley, "Speaker Identification and Verification Combined with Speaker Independent Word Recognition," Proc. 1981 Int'l Conf. Acoustics, Speech and Signal Processing, April 1981.

19) S.E.Levinson and K.L.Shipley, "A Conversational-Mode Airline Information and reservation System Using Speech input and Output," Bell System technical J., Vol. 59, No. 1. Jan. 1980, pp. 119-137

20) H.Sakoe, "Two-Level DP-Matching-A Dynamic Programming- Based Pattern Matching Algorithm for Connected Word Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-27, No. 6, Dec. 1979, pp. 558-595.

21) C.S.Myers and L.R.Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-29, 1981, pp. 285-296.

22) D.P.Huttenlocher and V.W.Zue, "A Model of lexical Access from Partial Phonetic Information," IEEE Int'l Conf. Acoustics, Speech and Signal Processing, Vol. 2, March, 1984, pp. 26.4.1-26.4.4.